

Metagenomic Bacterial Finishing at the JGI

Stephen Lowry, Alicia Clum, Eugene Goltsman, Hector Garcia Martin, Phil Hugenholtz, Alla Lapidus

US DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598



Abstract:

Metagenomic Bacterial Finishing at JGI.

Stephen Lowry, Alicia Clum, Eugene Goltsman, Hector Garcia Martin, Phil Hugenholtz, Alla Lapidus

DOE Joint Genome Institute, Walnut Creek, CA

The sequencing to completion of uncultured bacterial genomes from naturalistic communities is a demanding process in the best of cases. The complexity of the community, the quantity of genomic DNA available, the fraction of the total DNA collected representing the organism under study are all added to the normal difficulties of establishing a complete sequence.

At the JGI we have managed to complete the sequence of three metagenomic organisms, and are investigating a fourth, presenting a considerable range of difficulty.

Candidatus Korarchaeum cryptofilum OPF8, (NC_010482; GI170289627), is the first of this apparently ancient hyperthermophilic phyletic group to be sequenced (3). The ability to obtain ample DNA of near-monocultural purity and low strain complexity made this the most straightforward sort of metagenomic subject. The target organism constitutes ~40% of the Yellowstone thermal Obsidian Pool community. The community could be maintained in culture, and it was found that *K. cryptofilum* was the most resistant member to SDS lysis, thus allowing DNA purification to better than 90%. Its strain complexity was low as indicated by a SNP rate of ~0.2%.

Some organisms have remote or difficult habitats limiting the availability of source material. This is the case with the thermophile *Desulforudis audaxviator*, (NC_010424; GI169830219), from fractures in the earth's crust at a depth of 2800 meters in a South African gold mine (4).

The bacteria were collected on filters through which large amounts of subterranean water was passed. The surprising fact that this ecosystem contained but one species fortunately meant that the DNA yield of this one-time-only collection was sufficient to complete the genome.

Considerably more complex situations are the rule, as illustrated by the case of *Candidatus Accumulibacter phosphatis* Type IIA str. CU-1. This and closely related species are the principal actors in the sequestration of inorganic phosphate as intracellular polyphosphate in wastewater treatment facilities. Bioreactor sludge derived from a working facility in Wisconsin is a physically unresolvable mixture of organisms, with *A. phosphatis* predominating at about 40%. The entire DNA sample was sequenced and the resulting data then subjected to phylogenetic parsing using the PhyloPythia (2) binning technique, greatly reducing the complexity of the subclone libraries.

A single 5Mb chromosome was successfully sequenced, along with 3 plasmids of 167, 42 and 38kb. We are now approaching a still more difficult genome. *Candidatus Endomicrobium trichonymphae* is an intracellular symbiont of a flagellate protist, itself part of the normal hindgut community of a termite host. It is of interest in the pursuit of the efficient breakdown of cellulose and lignin necessary in the hoped-for use of bulk plant materials as CO₂-neutral fuel stocks.

Again partitioning with PhyloPythia was absolutely necessary. So far this has yielded some 42,000 Sanger reads. Of these about 40% of the assembled contigs are similar to related organisms. Also contigs from 454 pyrosequencing contributed about 750,000 bp of additional coverage (~0.5x). Additional difficulty arises from sample size and strain complexity. There are several closely related organisms with substantial representation. There are several hindguts in the sample, and the lines of descent may be relatively independent even at the protist level.

(1) Martin, H.G., Hugenholtz, P., et al. Nature Biotechnology, v24, no.10, 1263-69.

2006. Metagenomic Analysis of two Enhanced Biological Phosphorus Removal (EBPR) sludge communities

(2) McHardy et al. Appl. Environ Microbiol 2000, 66(3):1175-1182, Accurate phylogenetic classification of variable-length DNA fragments.

(3) James G. Elkins, et al., PNAS 2008 (in Press), The Korarchaeota: Archaeal

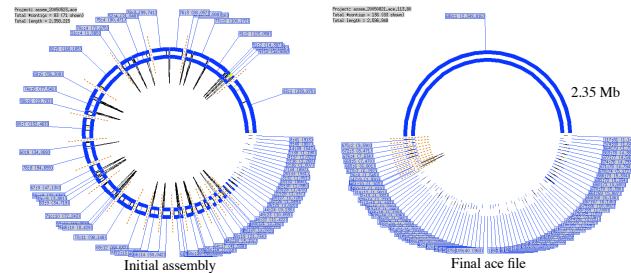
orphans representing an ancestral lineage of life

(4) Dylan Chivian, et al., Science 2008 (in Press), Environmental genomics reveals a single species ecosystem deep within the Earth.

Desulforudis audaxviator

Described as a 'one organism ecosystem' (4), which is to say it has been environmentally enriched. The representations in Orchid (available from the Stanford Genome Center) at right show clearly the low complexity of this community. (Orchid displays an ace file distributed around a circle, the heavy blue sections being continuous sequence.)

The left hand image is a representation of the contigs in the earliest assembly ace-file, while the right is of the final. Only about 100 reads remain unincorporated out of 28000.

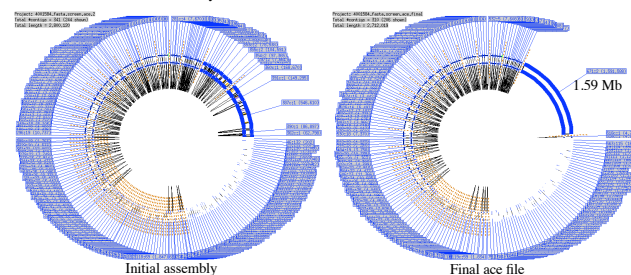


Korarchaeum cryptofilum OPF8

From a more complex thermophilic community, it was maintained as a community in the lab, and enriched by differential lysis. This organism was finished by conventional means. Its small size was helpful.

Though there was considerable enrichment, there is a great deal more representation of other species in this project than with *D. audaxviator*.

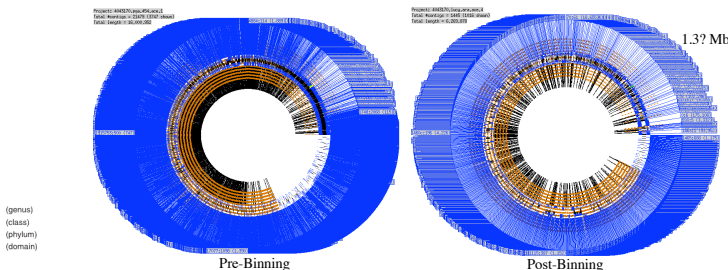
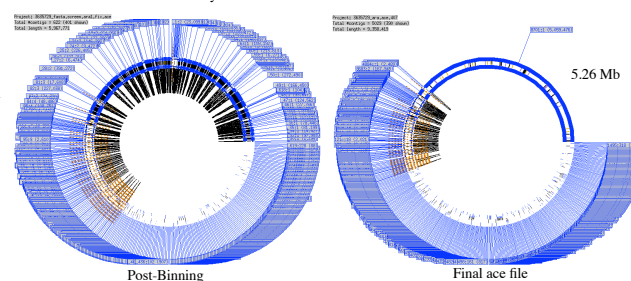
These first two organisms certainly represent the unusually straightforward end of the community finishing spectrum.



Accumulibacter phosphatis Type IIA str. CU-1

A more difficult, and undoubtedly more typical situation was presented by *A. phosphatis*. It was the first community subjected to PhyloPythia binning, reducing the number of reads by more than half and which eventually produced the assembly illustrated at right. One can see that after completion about 45% of the initial reads have fused into the final consensus.

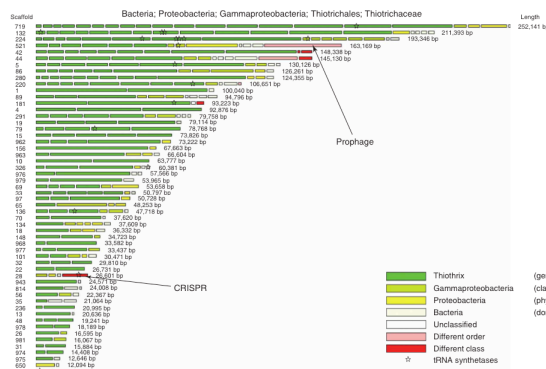
There is moderate polymorphism. It was our good fortune that 454 sequencing came on the scene as we considered finishing the genome. It allowed the bridging of many small gaps, and confirmed other only moderately supported regions.



Endomicrobium trichonymphae

A more challenging situation exists with *E. trichonymphae*. The sequence obtained from the combined hind-guts of a group of termite nest-mates was binned targeting a bacterial symbiote of a symbiotic flagellate protist. The enrichment reduced the read number by about 60%. The pertinent portion of the assembly at this point seems to consist of a number of strains considerably more distant from each other than the polymorphic members of the previous projects.

This can be illustrated by the cartoon on the right. In it, contigs with clear BLAST related sequence to a very closely related organism are aligned against that genome (AP009510.1), which is represented by the left vertical bar. The coloring is arbitrary to make contigs distinct.



PhyloPythia binning of Thiobacterales species (from McHardy et al (2))

Scaffolds which consist of mate-pair linked contigs, 62% of which were binned by PhyloPythia at the level of genus and 32% more at phylum level, with only 3% unclassified. Thiobacterales constituted the largest clade of gammaproteobacteria in an Australian waste treatment sludge community studied for its Accumulibacter species in parallel with the US species sequenced at the JGI.

organism	project #	enrich. mode	degree of enr.	GC	Mb	polym	454	subclone reads
<i>D. audaxviator</i>	4000602	environ.	100%	62%	2.4	low	Y	28167
<i>K. cryptofilum</i>	4001584	pchem	90%	49%	1.6	low	N	19688
<i>A. phosphatis</i>	3635729	binning	40%	57%	5.3	mod	Y	60579
<i>E. trichonymphae</i>	4043170	binning	< 40%	38%	1.37	high	Y	37300

There are certainly limits to the genomic resolution of communities, but luck and ingenuity have not been exhausted yet.